

BAKER & BOTTS, L.L.P.  
30 ROCKEFELLER PLAZA  
NEW YORK, NEW YORK 10112

TO ALL WHOM IT MAY CONCERN:

Be it known that WE, ANDREY RZHETSKY and SERGEY KALACHIKOV, citizens of Russia, whose post office addresses are 560 Riverside Drive, 11F New York, New York 10027; and 154 Haven Avenue, 1303, New York, New York 10032 respectively; MICHAEL O. KRAUTHAMMER, citizen of Switzerland, whose post office address is 27 W. 76th Street, Apt. 3A, New York, N.Y., 10023; CAROL FRIEDMAN and PAULINE KRA, citizens of the United States, whose post office addresses are 14 Dimitri Place, Larchmont, New York, 10538 and 109-14 Ascan Ave. Forest Hills, N.Y., 11375, respectively, have invented an improvement in

GENE DISCOVERY THROUGH COMPARISONS OF NETWORKS  
OF STRUCTURAL AND FUNCTIONAL RELATIONSHIPS  
AMONG KNOWN GENES AND PROTEINS

of which the following is a

SPECIFICATION

*ins  
diagram*

The invention described herein was funded in part by a grant from the National Library of Medicine, namely, Grant Number's LM06274 and LM05627. The United States Government may have certain rights to the invention. The present specification contains a computer program listing which appears as a microfiche Appendix H.

## STATEMENT REGARDING MATERIAL SUBJECT TO COPYRIGHT

A portion of the disclosure of this patent document contains material which is subject to copyright protection. The copyright owner has no objection to the facsimile reproduction by anyone of any portion of the patent document, as it appears in any patent granted from the present application or in the Patent and Trademark Office file or records available to the public, but otherwise reserves all copyright rights whatsoever.

An appendix containing source code listing utilized in practicing an exemplary embodiment of the invention is included as part of the Specification.

### 1. INTRODUCTION

The present invention relates to methods for identifying novel genes comprising: (i) generating one or more specialized databases containing information on gene/protein structure, function and/or regulatory interactions; and (ii) searching the specialized databases for homology or for a particular motif and thereby identifying a putative novel gene of interest. The invention may further comprise performing simulation and hypothesis testing to identify or confirm that the putative gene is a novel gene of interest.

The present invention relates to natural language processing and extraction of relational information associated with genes and proteins that are found in genomics journal articles. To enable access to information in textual form, the natural language processing system of the present invention provides a method for extracting and structuring information found in the literature in a form appropriate for subsequent applications. Specifically, the present invention provides for the generation of

specialized databases containing information on gene/protein structure, function and regulatory interactions based on the retrieval of such information from research articles and databases, and computer representation of such information in a manner that allows efficient access to the extracted information.

5                   The invention further provides for the use of the specialized databases for identifying novel genes based on detection of sequence similarities and domain/motif matches between genes/proteins, computation and interpretation of phylogenetic trees for multigene families, and analysis of homologous regulatory networks. The methods of the invention are based on the observation that functionally similar regulatory systems are  
10 generated during evolution by genetic duplication of ancestral genes. Thus, a comparison of homologous/similar networks within the same organism and between different species will allow the identification of genes absent in one of the systems under comparison. In this way genes that contribute to the phenotype of a specific disease associated with a particular biological system under analysis may be identified.

## 15                   2. BACKGROUND OF THE INVENTION

### 2.1. NATURAL LANGUAGE PROCESSING

Researchers working in molecular biology must constantly consider the information present in the literature relating to their regulatory systems of interest and the genes and proteins that operate within those systems. Unfortunately, to remain up-to-date  
20 on the relevant literature, the researcher is required to perform laborious reading and manual integration of research articles, each of which may address a narrow subject. Therefore, technology that enables rapid retrieval of information from literature and

manipulation of derived functional data should have a dramatic effect on the access of the researcher to important facts and ultimately should facilitate the discovery of novel human genes.

Natural language processing is an automated system that provides for a complex of programs for automatic retrieval of information from text analysis and for the computer representation of that information in a form that allows efficient access and extraction of that information. MedLee (Medical Language Extraction and Encoding System) has recently been successfully used for processing different types of medical texts as described in co-pending United States Patent Application Serial Number 09/370,329, incorporated herein in its entirety by reference (see also, Friedman et al., 1994, J. Amer. Med. Inf. Assoc. 1:161-174; Hripcsak et al. 1995, Ann. Intern. Med. 122:681-688; Hripcsak et al., 1998, Meth. Inform. Med.; Jain et al., 1996, Proc. AMIA Annu. Fall Symp. 542-546; Knirsch et al., 1998). When tested, MedLEE was on average as successful in retrieving reports associated with specified clinical connections as twelve medical experts invited for evaluation of the system.

Another text analysis technique has recently been developed that combines finite-state machines with statistical machine learning approaches. These models extract detailed semantic information from texts (e.g., see Hatzivassiloglou 1996, In Klavens, J.L., and Resnick, P.S. (eds) *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, MIT Press, Cambridge, MA) when extensive prior knowledge about the domain is not available. The techniques have been subsequently applied to the tasks of (i) automatically identifying medical terms for the automated summarization of

research articles reporting on clinical studies and (ii) sanitizing sensitive information in patient records so that they can be widely disseminated for research purposes.

A number of projects have also been developed as statistical information extraction tools that operate with limited or no prior knowledge about the application domain. These earlier efforts include XTRACT, a tool that recovers collocational restrictions between words that has been licensed to more than thirty sites worldwide (Smadja, F., 1993, J. Comp. Ling. 19:143-177), CHAMPOLLION, a system that retrieves bilingual mappings between words and phrases in parallel texts from different languages (Smadja, F. et al. 1996, J. Computational Linguistics 22:1-38), and a system that automatically aligns noisy, semi-parallel texts from different languages (Fung, P. and McKeown, K.R., 1997, Machine Translation 11:23-29).

## 2.2. IDENTIFICATION OF NOVEL GENES

A variety of different methods are currently utilized for the identification and characterization of novel genes. Perhaps the most widely used method for generating large quantities of sequence information is via high throughput nucleotide sequencing of random DNA fragments. A disadvantage associated with this gene discovery technique is that in most instances when genes are identified their function is unknown.

For identification of specific disease genes, positional cloning is currently the most widely used method. The positional cloning approach combines methods of formal genetics, physical mapping and mutation analysis and usually starts with a precise description of the disease phenotype and a tracing of the disease through families of affected individuals. Genetic linkage data obtained from the analysis of affected families

frequently allows the determination of an approximate genomic localization of the candidate disease gene with a precision of several millions of nucleotides. Once localized, the genetically defined chromosomal region is then recovered from genomic libraries as a contiguous set of genomic fragments. Genes residing in the disease-related region are determined by analysis of transcripts that are transcribed from the genomic fragment. From this analysis an initial set of candidate genes for a particular disease are identified based on the presence of the gene product in the biological system affected by disease and a correlation between its expression pattern and the pattern of disease progression.

10                    Important information for selection of candidate genes also comes from analysis of their homology with genes known to be part of the same or related biological system. Finally, the ultimate proof of association between a gene and a genetic disorder comes from mutational analysis of a gene in patients affected by the disorder and from demonstration of a statistical correlation between occurrence of mutation and the disease phenotype.

15                    Although positional cloning is a powerful method for gene discovery, the experimental method is extremely tedious and expensive. Moreover, disease genes implicated in genetically complex disorders, *i.e.*, those controlled by multiple loci, can hardly be found using this strategy because of the complications associated with multiple loci linkage analysis.

20                    Specialized databases for homology searches have also been utilized in disease gene discovery projects. In recent years a number of efficient sequence comparison tools have been developed such as the BLAST (Basic Local Alignment

Search Tool) family of programs designed for comparison of a single "search sequence" with a database (see Altschul et al., 1990, J. Mol. Biol. 215:403-410; Altschul et al., 1997, Nucleic Acids Res. 25:3389-3402), the family of Hidden Markov Model methods for comparison of a set of aligned sequences that usually represent a protein motif or domain with a database (e.g., Krogh et al., 1994, J. Mol. Biol. 235:1501-1531; Grundy et al., 1997, Biochem Biophys. Res. Commun. 231:760-6) and various other comparison tools (Wu et al., 1996, Comput. Appl. Biosci 12:109-118; Neuwald et al., 1995, Protein Sci. 4:1618-1632; Neuwald, 1997, Nucleic Acids Res. 25:1665-1677).

When used in disease gene discovery projects, homology searches can be enhanced by creating specialized databases that utilize statistical analysis for evaluating significance of sequence similarities in comparison of new sequences with a database of known sequence. Such databases are fine-tuned to the size of the database used (Altschul et al., 1990, J. Mol. Biol. 215:403-410; Altschul et al., 1997, Nucleic Acids Res. 25:3389-3402), so that the same level of homology between a search sequence and a database sequence can be determined to be highly significant if the search sequence is compared with a smaller database, or insignificant and thus undetectable, if the search sequence is compared with a larger database.

In alternatives to standard homology searches, in projects oriented towards gene discovery, researchers usually have some *a priori* knowledge about the set of genes/proteins that might display important similarity to the unknown new gene. Therefore, selecting an *a priori* defined set of genes/proteins for comparison with new experimental sequences is a feasible and useful strategy. This strategy was successfully

applied to search for homologs of disease genes in yeast and nematode genomes by Mushegian et al. (1997, Proc. Natl. Acad. Sci USA 94:5831-5836).

Two homologous genes taken from different species that originate from the nearest common ancestor by speciation are referred to as orthologs, while any two  
 5 genes that originate from a common ancestor via a series of events involving intragenomic duplications are called paralogs. Tatusov et al. (1994, Proc. Natl. Acad. Sci USA 91:12091-12095) describe comparisons of proteins encoded by the genomes of different phylogenetic lineages and elucidation of consistent patterns of sequence similarities permitting the delineation of clusters of orthologous groups (COGs). Each  
 10 COG consists of individual orthologous genes or orthologous groups of paralogs from different phylogenetic lineages. Since orthologs typically have the same function, the classification of known genes and proteins into clusters of orthologous groups permits the assignment of a function to a newly discovered gene or protein by merely classifying it into a COG. Although Tatusov describes a method for assigning a function to a newly  
 15 discovered gene, he does not describe a method for predicting the existence of undiscovered genes. In addition, Yuan, et al. attempted simultaneous reconstruction of a species tree and identification of paralogous groups of sequences and detection of orthologs in sequence databases (Yuan et al., 1998, *Bioinformatics* 143:285-289).

Other groups have aimed at capturing interactions among molecules  
 20 through the use of programs designed to compare structures and functions of proteins (Kazic 1994, In: Molecular Modeling: From Virtual Tools to Real Problems, Kumosinski, T. and Liebman, M.N. (Eds.), American Chemical Society, Washington, D.C. pp. 486-494; Kazic, 1994, In: New Data Challenges in Our Information Age



Glaesar, P.S. and Millward, M.T.L. (Eds.). Proceedings of the Thirteenth International CODATA Secretariat, Paris pp. C133-C140; Goto et al., 1997, Pac. Symp. Biocomput. p. 175-186; Bono et al., 1998, Genome Res. 8:203-210; Selkov et al., 1996, Nucleic Acids Res. 24:26-28). These projects are significantly different from the inventive methods described herein because they do not describe methods for deducing the existence of as yet unknown genes based on comparisons of regulatory pathways and gene structure between one or more species. The present invention provides a method for increasing the sensitivity of analysis methods through the generation of specialized databases.

### 3. SUMMARY OF THE INVENTION

In accordance with the present invention there is provided methods for identification of novel genes comprising (i) generating one or more specialized databases containing information on gene/protein structure, function and/or regulatory interactions; and (ii) searching the specialized databases for homology or for a particular motif and thereby identifying a putative novel gene of interest. The invention may further comprise performing simulation and hypothesis testing to identify or confirm that the putative gene is a novel gene of interest.

The invention is based, in part, on the observation that functionally similar regulatory systems are generated during evolution by genetic duplication of ancestral genes. Thus, by comparing phylogenetic trees or regulatory networks and identifying genes and/or proteins absent in one system under comparison, the existence of as yet unidentified genes and/or proteins can be predicted. To make meaningful comparisons of phylogenetic trees it is necessary to distinguish between orthologs and paralogs. The

present invention provides a method useful for discriminating between orthologs and paralogs and inferring the existence of as yet unidentified genes and/or proteins.

The present invention relates to natural language processing and extraction of relational information associated with genes and proteins that are found in genomics journal articles. Specifically, the natural language processing system of the invention is used to parse the articles published in biological journals focusing on structure and interactions among genes and proteins followed by computer representation of such interactions.

In accordance with the present invention, specialized databases are developed that contain information on gene/protein structure and interactions based on information derived from preexisting databases and/or research articles including information on interactions among genes and proteins, their domain/motif structure and their subcellular and tissue expression/distribution patterns.

The invention relates to a sequence analysis program which utilizes the specialized database for comparison of a single sequence, processing the output into a sequence alignment, computing phylogenetic trees, and analyzing these trees to predict undiscovered genes. This program also includes a set of tools for generating motif/domain models from multiple sequence alignments of known genes and for using these models for extraction of structurally and/or functionally homologous sequences from databases which contain raw sequence data.

The invention further provides for a simulation and hypothesis testing program which relies on the specialized databases of gene/protein interactions for identifying potentially undiscovered members of multigene families through comparisons

of regulatory networks for different species and testing hypotheses with regard to regulatory cascades. A comparison of homologous regulatory networks within the same organism and between different species of organisms will allow the identification of genes absent in one of the systems under comparison, thus providing a set of candidate  
5 genes. In this way, genes that contribute to the phenotype of a specific disease associated with a particular biological system under analysis may be identified, mapped and subjected to mutational analysis and functional studies.

#### 4. BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 is a block diagram illustrating the three major programs of the  
10 method according to the present invention: (i) the generation of specialized databases based on information on gene/protein structure, function and regulatory interactions derived from research papers and databases; (ii) sequence analysis; and (iii) simulation and hypothesis testing;

Figure 2 is a block diagram of an information extraction system in  
15 accordance with a preferred embodiment of the present invention;

Figure 3 is a diagram illustrating the object representation of molecules and relations between them;

Figure 4 shows a set of keywords defining proteins involved in apoptosis pathways, these keywords having been utilized for generating a specialized sequence  
20 database Apoptosis3, this list having been compiled manually for testing the concept of specialized databases;

Figure 5 shows a “species tree,” which is a graph depicting the correct order of speciation events leading to a set of present day species; a “gene tree,” which is a graph depicting a history of a few genes from the same species, where each species can be represented by multiple paralogous genes (because the set of known genes is incomplete for most genomes, and there are often multiple representations of the same gene family in the same genome, the gene tree can be drastically different from the corresponding species-tree); and a “reconciled tree”, which is the gene tree that would be obtained if gene deletions were completely forbidden and all genes were known for all species under analysis;

Figure 6 shows the original tree of ALDH sequences, indicating sequence clusters where bacterial, plant, fungal and nematode orthologous genes are present, but a human ortholog was not yet known;

Figure 7 shows the same phylogenetic tree as in Figure 6 with an additional human protein, referred to as antiquitin which was discovered by the method of the invention;

Figure 8 is a schematic diagram illustrating functional network-based gene discovery in accordance with the present invention;

Figure 9A presents diagrams depicting the regulatory relationships among hypothetical proteins (denoted with Arabic numerals) of hypothetical species A and B.

Proteins in different species denoted with the same numeral are considered orthologous. The diagrams show that regulatory relationships between a pair of proteins can be of three different kinds;

Figure 9B, 9C, and 9D are diagrams representing Boolean operations OR, AND, and XOR, on arcs of the two oriented graphs of Figure 9A, the same operations being applicable to the set of vertices of the two oriented graphs;

Figure 10 is a diagram representing a hypothetical example of defining homologous protein networks in two different species using protein motifs, the diagram showing only two hypothetical proteins (1 and 2) for species A and three hypothetical proteins (1, 3, and 4) for species B. Protein 1 in both species has motifs  $\alpha$  and  $\beta$ , protein 2 has motifs  $\delta$ ,  $\epsilon$ , and  $\zeta$ , and proteins 3 and 4 have motifs  $\delta$  and  $\zeta$ , and  $\epsilon$ , respectively. The motif analysis can indicate that proteins 3 and 4 in species B may collectively perform the same function as protein 2 in species A;

Figure 11A and 11B are diagrams respectively representing hypothetical examples of evaluating the impact of a "knockout" of hypothetical gene A on the expression of a hypothetical gene B. The effect of knock-out of gene A calculated by multiplication along the shortest pathway connecting genes A and B is inhibition of gene B, the resulting effect being zero if the orientation of only one arc in the same pathway is reversed;

Figure 12 is a flow chart representing the scheme of gene discovery analysis involving motif/domain analysis in accordance with the present invention; and

Figure 13 Identification of genes in *C. elegans* containing either POZ or kelch domains. The protein accession numbers are indicated adjacent to the different protein domains. The protein corresponding to accession number gi/1132541 contains a POZ domain, death domain, kinase domain and heat repeat.

Figure 14A. Two human sequences with the closest homology to the *C. elegans* sequence gi/1132541.

Figure 14B. Computed gene tree indicating that the identified human gene represents an ortholog of the *C. elegans* gene gi/1132541.

5 Figure 14C. Nucleotide sequence of the death domain gene.

Figure 14D. Deduced amino acid sequence of the death domain protein.

Figure 15. Identification of candidate gene implicated in the etiology of Chronic Lymphocytic Leukemia (CLL). Sequence homology between a CLL region open reading frame and mouse Rpt1 (sp/P15533/RPT1) is presented.

10 Figure 16A-B. Model of regulatory functions of Rpt1. Figure 16A indicates that in mouse T lymphocytes Rpt1 serves as a repressor of the gene for interleukin 2 receptor (IL-2R). Figure 16B demonstrates that when Rpt1 is knocked out, the regulatory effect is manifested as a block of the apoptotic pathway for T-lymphocytes resulting in accumulation of T-lymphocytes in blood.

15 Figure 17A. Two EST sequences identified by searching a protein dbEST using the mouse Mad3 protein as a query.

Figure 17B. Nucleotide sequence of the human Mad3 gene.

20 Figure 17C. Complete sequence of the human Mad3 protein. A search was conducted to identify overlapping sequences. The complete sequence of the gene was assembled and the amino acid sequence deduced. The translated human Mad3 sequence consists of 206 amino acid residues 81% of which are identical to the mouse Mad3 protein.

Figure 17D. Multiple alignment of the human Mad3 amino acid sequence with known Mad proteins.

Figure 18A. Phylogenetic tree indicating relationship between three known mouse Mad genes and their two human homologs.

5 Figure 18B. Phylogenetic tree including new human Mad3 sequence. The phylogenetic tree indicates that the new human gene belongs to the family of Mad proteins and is an ortholog of mouse Mad3.

## 5. DETAILED DESCRIPTION OF THE INVENTION

10 The present invention provides methods for identification of novel genes comprising: (i) generating specialized databases containing information on gene/protein structure, function and regulatory interactions and, (ii) sequence analysis which includes homology searches and motif analysis thereby identifying a putative novel gene of interest. The invention may further comprise performing simulation and hypothesis  
15 testing to identify or confirm that the putative gene is a novel gene of interest.

The specialized databases are constructed utilizing information concerning gene/protein structure or function derived from unpublished data, research articles and/or existing databases. The specialized databases can be used to identify novel genes by:

(i) searching for motif/domain combinations characteristic for a putative gene of interest;

20 (ii) phylogenetic tree analysis of homologous genes for predicting the existence of yet undiscovered genes; (iii) comparing members of interactive gene/protein networks from different species for predicting the existence of yet undiscovered genes; and (iv) testing a

hypothesis with regard to known interactions of homologs from other species in regulatory pathways.

### 5.1. THE NATURAL LANGUAGE PROCESSING

The present invention relates to a natural language processing system that is designed to parse the electronic versions of articles published in journals that report on structural interactions among genes and proteins. The system provides a method for extracting information on interactions among genes and proteins, their domain/motif structure, and/or their sub-cellular and tissue expression/distribution patterns, followed by computer representation of such information.

The general natural language-processing system of the invention is schematically depicted in Figure 2. The collection phase automatically collects articles from appropriate literature, and selects articles that contain relevant information using Keyword search techniques. In the next phase, the preprocessor standardizes the selected articles so that they consist of tagged ASCII text where the tags delineate critical components of the article. The next phase, termed the extraction phase, retrieves and classifies biological entities, *i.e.*, as names of proteins, genes and small molecules. In addition, the relationship extraction phase recovers structural relationships between the entities. This phase is followed by a phase which performs an analysis of the sequence of events.

The final phase of the system processes the output extracted from an article to remove redundancies, inconsistencies and to incorporate implicit information before adding the extracted knowledge consisting of biological entities, their attributes,



conditional constraints, and relationships between them, for subsequent use in analysis and hypothesis testing. The information extraction system as depicted in Figure 2, referred to herein as "GENIE," is designed for use as a general processor within the domain of genomics literature although the system may also be used in other specialized domains. GENIE is an adaptation of MedLEE developed for the medical domain. GENIE uses the same source code as MedLEE but the Lexicons and grammar were adapted for genomics literature.

The information extraction system of the present invention is described below, by way of example, with reference to the genomics domain uses of GENIE. It is written in Quintus Prolog and uses the Unix or Windows operating systems, as described in detail below.

A natural-language phrase included in text document is understood as a delimited string comprising natural-language terms or words. The string is computer readable as obtained, *e.g.*, from a pre-existing database, a keyboard input, optical scanning of typed or handwritten text, or processed voice input. The delimiter may be a period, a semicolon, an end-of-message signal, a new-paragraph signal, or any other suitable symbol recognizable for this purpose. Within the phrase, the terms may be separated by another type of delimiter such as a blank or another suitable symbol.

As a result of phrase parsing, terms in a natural-language phrase are classified, (*e.g.*, as referring to a gene, a protein, or their interactions) and the relationships between the interactions are established and represented in a standard form. For example, in the sentence "Rap inhibited fyn", the structured form would be:

[action,inactivate,[protein,rap],[protein,fyn]].

In such an example, the interaction is “inactivate”, the agent is “Rap” and the target is “fyn.” More complex sentences consisting of nested relationships, such as “The

activation of BAD was suppressed by the phosphorylation of JNK” can also be parsed

5 and represented appropriately. The structured output form for this sentence would be:

[action,inactivate,[action,phosphorylate,x,[protein,jnk],[action,activate,x,[protein,bad]]]

In the first example, the primary interaction is “inactivate”; in the second example, an

interaction “phosphorylate” is the agent where the protein “jnk” is its target (the agent of

“phosphorylate” is not specified and thus is represented as “x”). In this example, the

10 target of “inactivate” is also an interaction “activate” where the target is the protein “bad” and the agent is unknown.

While parsing is based on both syntactic and semantic grammatical patterns, the substances in a domain are normally only semantic categories such as

“protein”, “gene”, and “small molecule.” There are no corresponding syntactic

15 categories needed for these substances because they are normally all nouns. However, each action can be categorized both semantically and syntactically. An action, which is a semantic category, can generally occur syntactically as a verb “inactivate” or as a noun “inactivation.” Therefore there are two sets of lexical entries for the actions: syntactic

and semantic. The syntactic lexicon for actions specifies the main syntactic category such as “v” for verb, “ving” for progressive form of verb, and “activation” for noun.

The semantic entries for actions not only categorize the actions, but also specify features for each action. For example, one feature provides the number of arguments that are

5 expected for the action, *i.e.*, some actions are associated with two arguments because they have an agent and a target as “inactivate”, and others just have an agent “mutate.” The lexicon of substances and structures appears as Appendix A; the syntactic lexicon for actions appears as Appendix B; and the semantic lexicon of actions appears as Appendix C.

10 A second feature specifies whether or not the arguments should be reversed when obtaining the target form. For example the arguments of “attributable to” should be reversed, *i.e.*, in “the phosphorylation of jnk is attributable to the activation of bad”, the underlying action is “cause” (from “attributable to”), the agent is the “activation of bad” and the target is “the phoshorylation of jnk”), whereas the arguments of  
15 “activates” is not( *i.e.* in “jnk activates bad” , the agent is “jnk” and the target is “bad”).

Figure 2 shows a preprocessor module of GENIE by which natural-language input text is received. The preprocessor thus performs lexical lookup to identify and categorize multi-word and single word phases within each sentence. The output of this component consists of a list of word elements where each element is associated with

a word or multi-word phrase in the report. For example, assuming that the sentence “bad functions as a negative regulator of the activation of jnk” is at the beginning of the report, it would be represented as a list of elements where each element is a word or phrase. For example, element 1 is associated with “bad”, element 2 with the multi-word phrase “functions as a negative regulator of”, element 8 with “the”, and element 9 with “activation”. The remainder of the list of word positions would be associated with the remaining words in the report. Some of the phrases may not need lexical lookup because they already have been tagged by a previous component. Such a tagging system is described below in Section 5.2.

The second component of the GENIE system is the parser. It utilizes the grammar and categories assigned to the phrases of a sentence to recognize well-formed syntactic and semantic patterns in the sentence and to generate structured output forms. The parser proceeds by starting at the beginning of the sentence element list and following the grammar rules. When a semantic or syntactic category is reached in the grammar, the lexical item corresponding to the next available unmatched element is obtained and its corresponding lexical definition is checked to see whether or not it matches the grammar category. If it does match, the word or phrase is removed from the unmatched sentence list, and the parsing proceeds. If a match is not obtained, an alternative grammar rule is tried. If no analysis can be obtained, an error recovery

procedure is followed so that a partial analysis is attempted. The actual grammar used for GENIE appears as Appendix D.

The parser module of GENIE uses the lexicon, and a grammar module to generate target forms. Thus, in addition to parsing of complete phrases, subphrase  
 5 parsing can be used to an advantage where highest accuracy is not required. In case a phrase cannot be parsed in its entirety, one or several attempts can be made to parse a portion of the phrase for obtaining useful information in spite of a possible loss of information.

Conveniently, each module is software-implemented and stored in  
 10 random-access memory of a suitable computer, *e.g.*, a work-station computer. The software can be in the form of executable object code, obtained, *e.g.*, by compiling from source code. Source code interpretation is not precluded. Source code can be in the form of sequence-controlled instructions as in Fortran, Pascal or "C", for example.  
 Alternatively, a rule-based system can be used such a Prolog, where suitable sequencing  
 15 is chosen by the system at run-time.

An illustrative portion of the GENIE system is shown in the Appendix D in the form of a Prolog source listing with comments. The following is further to the comments.

*Process\_sents* with *get\_inputsents*, *process\_sects* and *outputresults* reads in an input stream, processes sections of the input stream according to parameter settings, and produces output according to the settings, respectively. Among parameters supplied to *Process\_sents* are the following: Mode (specifying the parsing mode) and Protocol (html or plain). *Process\_sents* is called by another predicate, after user-specified parameters have been processed.

The parsing modes are selected by GENIE so as to parse a sentence or phrase structure using a grammar that includes one or more patterns of semantic and syntactic categories that are well-formed. For example, for the phrase “bad inactivates jnk”, a legitimate pattern can be substance1 action substance2, wherein substance1 = protein bad, action = “inactivates” and substance2 = “jnk.” However, if parsing fails, various error recovery modes are utilized in order to achieve robustness. The error recovery techniques use methods such as segmenting the sentence, processing large chunks of the sentence, and processing local phrases. Each recovery technique is likely to increase sensitivity but decrease specificity and precision. Sensitivity is the performance measure equal to the true positive rate of the natural language processing, *i.e.*, the ratio of information extracted by the natural language processing system that should have been extracted. Specificity is the performance measure equal to the true negative information rate of the system, *i.e.*, the ratio of information not extracted by the

NLP system that should not have been extracted. Precision is the reliability of the system, *i.e.*, the ratio of information extracted correctly compared to all the information that was extracted. In processing a report, the most specific mode is attempted first, and successive less specific modes are used only if needed.

5                   In accordance with the preferred embodiments of the present invention, the parser of Figure 2 includes five parsing modes, Modes 1 through 5, for parsing sentences or phrases. Nominally, the parser is configured to first select Mode 1. If Mode 1 is not possible, the program continues with Mode 2 and so forth until parsing is complete. With Mode 1, the initial segment is the entire sentence and all words in the  
10                   segment must be defined. This mode requires a well-formed pattern for the complete segment.

                  Mode 2 requires that the sentence or phrase be segmented at certain types of words or phrases, *e.g.*, “is attributable to.” Here, an attempt is made to recognize each segment independently, *i.e.*, a first segment ending with the word “is” and a second  
15                   segment beginning with the word after “to.” The segmenting process is repeated until an analysis of each segment is obtained or until segmenting is no longer possible.

                  Mode 3 requires a well-formed pattern for the “largest” prefix of the segment, *i.e.*, usually at the beginning of the segment. This occurs when a sentence contains a pattern at the end which is not in the grammar but a beginning portion that is

included. For example, in “bad inactivates jnk at this time”, the beginning of the sentence “bad inactivates jnk” will be parsed and the remainder will be skipped.

Mode 4 requires that undefined words be skipped and an analysis be attempted in accordance with Mode 1. Mode 4 is useful where there are typographical errors and unknown words. For example, in the phrase “abc bad inactivates jnk”, the word *abc* is unknown to the system and will be ignored but the remainder of the phrase will be parsed.

Mode 5 first requires that the first word or phrase in the segment associated with an action be found. Next, an attempt is made to recognize the phrase starting with the leftmost recognizable argument. For example, in “during bad inactivates jnk on the fifth day,” the phrase “bad inactivates jnk” will be parsed and the remaining words will not be. If no analysis is found, recognition is retried at the next possible argument to the right. This process continues until an analysis is found.

*Process\_sects* with *get\_section* and *parse\_sentences* gets each section and generates intermediate output for the sentences in each section.

*Write* produces the output as a list consisting of relations and interactions

*Setargs* sets arguments or parameter values based on user input or by default.



The structured output generated by the GENIE program uses a frame-based representation. Each frame specifies the informational type, the value, and arguments or modifier slots which are also frames. Consider the text data input “bad inactivates the phosphorylation of jnk.” A corresponding output, as shown below, is a

5 frame denoting an action, which has the value inactivate; in addition, there are two arguments. The first argument is a protein bad and the second argument is an action with the value phosphorylate, which has two arguments. The first argument is x signifying that the agent has not been specified; the second argument is a protein with the value jnk. The second argument is the target:

10 [action,inactive,[protein,bad],[action,phosphorylate,x,[protein,jnk

In summary, a computer system has been disclosed that generates structured information concerning protein and gene interactions and relationships.

## 5.2. USE OF BLAST FOR FINDING GENE AND PROTEIN NAMES IN JOURNAL ARTICLES

15 In a specific embodiment of the invention, an exhaustive list of gene and protein names, extracted from GeneBank, is translated into a different alphabet system by substituting each character in the name with a predetermined unique nucleotide combination. The encoded names are then imported into the BLAST database using the

FASTA format. The scientific journals are translated, using the same nucleotide combinations, into a continuous string of nucleotides. A query is then used to match the translated journals against the nucleotide representation of gene and protein names in the BLAST database. Significant alignments associated with gene and protein names are listed in the BLAST output file, which is subsequently processed using Perl-scripts. The final result consists of the original journal article with XML tags surrounding the gene and protein names.

To adapt the problem to BLAST's statistical foundation, different measures were undertaken to limit the output to the most relevant gene and protein names. In addition, in order to fine-tune the matching process, different BLAST parameters were adjusted, such as the *word size* (which sets the size of the high scoring words, thus influencing the sensitivity of finding HSPs) and *mismatch penalty* (exact vs approximate matching).

In a specific embodiment of the invention, gene and protein names are extracted from GeneBank's gene symbol index file. The following is an excerpt of the file after discarding entries that are either composed of only numbers or of less than two alphabetic letters:

gfap gamma  
hox a10  
hox a1

wac 3'-end  
 pit-1/ghf-1 variant  
 [...]

Sub 5  
 B2

This list of gene and protein names is translated into a different alphabet system by substituting each character in the name with a predetermined unique nucleotide combination. The conversion chart is listed in Appendix E. The encoded names are then imported into the BLAST database using the FASTA format. For example, the first entry in the list above is "gfap gamma." After translation using the conversion chart, the same name appears as follows:

10 AGCAACTAAACACCCATCCAAGCAAACACACACACAAAC

Thus, the complete FASTA entry looks like this:

>gi|1 species,gp,gfap gamma

AAGCAACTAAACACCCATCCAAGCAAACACACACACAAAC

In FASTA, the definition line (marked with '>') contains information about the database entry. This line can contain any kind of information. The information important for this particular example is the third entry in the definition line, 'gp', that specifies that the name can represent a gene *or* a protein. If the name is unambiguous, then the definition line states that the name is only associated with a gene ('g') or protein ('p'). The fourth entry in the definition line is the name of the protein or gene, "gfap gamma" in this case.

15

20

The second line in the FASTA format normally contains the actual sequence of the protein/gene. In the example presented, the second line contains the translated protein or gene name.

- 5 All gene and protein names are translated into the nucleotide representation and converted into the FASTA format. Then, the database containing these FASTA entries are specially compiled for use in BLAST queries using a program that is included in the BLAST package called "formatdb".

10 Thus, the scientific journals are translated, using the same nucleotide combinations, into a continuous string of nucleotides. For example, the sentence "In the absence of costimulation, T cells activated through their antigen ..." is translated into

"AAGTACAGATCCACGGAAGGAACGATCCAAACAAAGACGCAACGACAGAA  
ATAACGATCCACATAACTATCCAAATACATACGCACGGAAGTACACACGTAA  
TTAAACACGGAAGTACATACAGATCCATCCACGGATCCAAATAACGAATTAA  
TTACGCATCCAAACAAATACGGAAGTACTCAAACACGGAACGAACCATCCAC  
15 GGAAGGACCTACATACGTAAGCAAGGATCCACGGAAGGAACGAAGTACCTA  
TCCAAACACAGACGGAAGTAAGCAACGACAGATCC "

A query is then used to match the translated journals against the nucleotide representation of gene and protein names in the BLAST database. The query is executed

using the blastall program that is included in the BLAST package. The query line looks like:

```
blastall -p blastn -d FASTA.dat -i query.txt
```

The flag 'p' denotes the sub-program (blastn is a sub-program of blastall that performs nucleotide matches), 'd' denotes the file that contains the FASTA entries and 'i' denotes the translated query text.

Significant alignments associated with gene and protein names are listed in the BLAST output file. This is an excerpt from a BLAST output file:

```

10      gi|63624 species,gp,ner
      Length = 12
      Score = 24.4 bits (12), Expect = 3e-05
      Identities = 12/12 (100%)
      Strand = Plus / Plus
15      Query: 729   acagaacgacct 740
      Sbjct: 1     acagaacgacct 12
```

The first line denotes the database entry. The second line denotes the database sequence length, followed by the alignment score and the E-value. The next line indicates paired matches, mismatches and gapped alignment (the latter two are not shown in this example). The lines 'Query' and 'Sbjct' show the actual alignment between the query and database sequence. This output file is subsequently processed using a Perl-script (see Appendix F). The script shown in Appendix G scans the output file, which is

sometimes several megabytes long, for any segments that start at position 1 of the database sequence (thus disregarding any segments that are only part of the sequence). In addition, the script allows for 10% mismatches between the aligned sequences for long sequences (as shown in the script of Appendix E), or 0% mismatches for short sequences.

- 5 After scanning the output file, an intermediary file that lists the candidate sequences is created:

	tran 365 381 gp 18493
	tran 1 17 gp 18493
	peci 549 565 gp 58106
10	il-2 621 637 gp 82396
	il-2 325 341 gp 82396
	gati 193 209 gp 92088
	prod 641 657 gp 52292
	rap1 105 121 gp 49898
15	spec 545 561 gp 33183
	crip 385 401 gp 118905
	crip 21 37 gp 118905
	as 161 177 gp 133961
	her 65 77 gp 88411

- 20 The intermediary file lists the name of the sequence, followed by the starting and end point in the query sequence (corresponds to where the two sequences matched), the semantic class of the name (protein, gene or protein/gene). The last number is not considered.

- The intermediary file is then scanned by another Perl program (Appendix  
25 G). This program compares the starting end points with the actual text, making sure that

the matched name is an 'autonomous' entity in the query text. For example, while "per" in " per gene" should be recognized as a gene name, "per" in "personal" should not be recognized as a gene name. The program recognizes other characters than the space character delimiting an 'autonomous' gene or protein name. In addition, the script looks for plurals of words. For example, " interleukins" should be recognized as a protein name, although only the singular form, "interleukin", is in the database.

The final result consists of the original journal article with XML tags surrounding the gene and protein names. This is done using the same script as in

#### Appendix G:

blocked <phr sem="gp">T cell antigen receptor</phr> (TCR)- and <phr sem="gp">CD28</phr>-mediated <phr sem="gp">IL-2</phr> gene transcription. Therefore, <phr sem="gp">Rap1</phr> functions as a negative regulator of...

To adapt the problem to BLAST's statistical foundation, different measures were undertaken to limit the output to the most relevant gene and protein names.

BLAST is sensitive to the search space the program works in. Thus, given a long query sequence and a large sequence database, matches have a lower statistical significance because the chances are higher that the matches could have occurred by chance alone. In addition, matches with few letters have a lower statistical significance

than matches with many letters. In order to find all true matches with any significance level, some measures were undertaken to address this problem. For example, (i) the query sequence was divided into 10 equal length parts, *i.e.*, the journal article was divided into 10 parts and 10 different queries are run on each part separately; (ii) the sequence database (with the gene and protein names) is separated into 5 databases, each containing protein/gene names of different length; (iii) gene and protein names with less than 3 letters in the database were 'expanded', *i.e.*, spaces were added at the beginning and the end of the name. Doing so, the statistical significance of a match containing a short name was higher. A space does not only include an empty character. For example, a gene name "k4" could occur in a journal article as "kinin 4 (k4)". It was therefore important to define several characters as substitutes for a space character. The alphabet in Appendix E defines the nucleotide combination ATCC as such a substitute.

Working with nucleotides implies that errors involving reading frames must be addressed. For example, working with a code of four letters, the nucleotide combination ATCTGTCACG could mean ATCT/GTCA or TCTG/TCAC or CTGT/CACG. Since the text is translated into a nucleotide combination, only one of these possibilities is correct. But BLAST can not distinguish between these solutions, *i.e.*, BLAST would potentially match a database sequence to a wrong reading frame in the



query sequence, producing many nonsense results that could compromise the significance of true results.

The solution to this problem is a comma-free code. A comma free code knows only one correct reading frame. BLAST therefore does not produce any nonsense results. A comma-free code consists of only one permutation of a nucleotide combination. For example, given the nucleotide combination ATCC and its permutations CATC, CCAT and TCCA, only ONE of these permutations would be included in a comma-free code. The code in Appendix E does represent a comma free code. Comma-free codes were discussed in the early days of DNA research (Crick et al., Proc. Natl. Acad. Sci. 43:416-421).

In order to fine-tune the matching process, different BLAST parameters must be adjusted, for example: *word size* (which sets the size of the high scoring words, thus influencing the sensitivity of finding HSPs); *mismatch penalty* (exact vs approximate matching); *numbers of alignments to show* (true matches of low significance can sometimes be at the very end of the BLAST output, therefore many alignments have to be shown); and *expectation value* (which sets the significance value for matches in the output file).

### 5.3. GENERATION OF SPECIALIZED DATABASES

In accordance with the present invention, specialized databases may be developed that contain information derived from unpublished data, publications such as research articles, theses, posters, abstracts, etc. and/or databases concerning interactions among genes and proteins, their domain/motif structure, and their biological functions.

For example, but not by way of limitation, a specialized database may be prepared as follows. Protein and gene sequences may be provided, for example, by the Java program PsiRetrieve which allows for quick retrieval of protein or nucleotide sequences from binary BLAST databases by sequence accession number, keyword or groups of keywords, or species name. In addition, using the program PsiRetriever, sequences encoding the proteins of interest may be retrieved from the non-redundant (NCBI) database of protein sequences and stored as a FASTA file. The FASTA file is then converted into a binary blast database using the program FORMATDB from the BLAST suit of programs.

Known motifs/domains for proteins may also be collected using the flat file versions of major protein databases, such as SwissProt (<http://expasy.hcage.ch/sprot>) and the non-redundant database of NCBI (<http://www3.ncbi.nlm.nih.gov>). The databases can be downloaded and searched for the keywords "motif" and "domain" in the feature tables of proteins. In addition, existing databases of motifs and domains, such as

BLOCKS (<http://dupsas.Weizmann.ac.il/bcd/bcdparent//databanksblocks/hfml>) and pfam(<http://www.sanger.ac.uk//software/pfam>; <http://pfm.wustl.edu>), can be downloaded (Henikoff et al., 1991, NAR 19:6565-6572). Still further, it is understood that any publically available database containing gene/protein sequences may be utilized to generate the specialized databases for use in the practice of the present invention.

Homologous sequences may be aligned using, for example, the CLUSTALW program (Higgins, et al. 1996 Methods in Enzymology 266: 383-402). A protein's sequence corresponding to each domain/motif can be identified, saved and used for building a Hidden Markov Model (HMM) of the domain/motif using a HMMER and HMMER2 packages (see, Durbin, R. et al. 1998 in Biological Sequence Analysis: Probablistic Models of Proteins and Nucleic Acids). HMMER and HMMER2 packages are useful for (i) building HMMs from sets of aligned protein or nucleotide sequences, and (ii) comparing the HMMs with sequence databases aimed at identifying significant similarities of HMMs with database sequences. Both nucleotide and protein databases can be used for this purpose. Alternatives to the Hidden Markov Model method for building domain/motif models include neural network motif analysis (Wu, C.H. et al., 1996, Comput Appl Biosci 12, 109-18; Hirst, J.D., 1991, Protein Eng 4:615-23) and positional weight matrix analysis (Claverie, J.M., 1994, Comput Chem 18:287-94;

Venezia, D., 1993, Comput Appl Biosci 9:65-9; Bucher, P. 1996, Comput Chem, 20:3-23; Tatusov, R.L., 1994, Proc Natl Acad Sci USA 91:12091-5).

Once a comprehensive collection of motifs/domains is created, each particular protein may be compared against a complete database of HMMs to identify known motifs and domains.

The Hidden Markov Model (HMM) is built using the following steps:

- A1. Start with a motif/domain name and a single amino acid sequence representing a domain or motif.
- A2. Do PSI-BLAST (BLASTPGP) search with the motif/domain sequence against a protein non-redundant database.
- A3. Retrieve the sequences identified in the database search from the protein sequence database. Exclude low-complexity sequences, short or incomplete sequences and sequences with similarity score above a selected threshold of PPD value  $<0.001$
- A4. Align the set of sequences with CLUSTALW (or other multiple sequence alignment program).
- A5. Use the set of aligned sequences for building HMM with the programs provided with HMMER and HMMER2 packages (see Hughey and Krogh 1996, J. Mol. Biol. 235:1501-1531).

- A6. Do a new database search comparing new HMM with the non-redundant protein database.
- A7. Continue steps A3-A6 until the convergence of the Markov model *i.e.*, until no new sequences are identified, or the maximum allowed number of iterations as defined by the user is reached. (Hugh R. and Krogh A., 1996, Comput. Appl. Biosci. 12: 95-107).

In addition, in yet another embodiment of the invention, a specialized database may be designed to contain a semantic model of proteins and of the possible interactions between them. Such databases are particularly useful for computation and analysis of regulatory networks between proteins. The semantic model is designed for representing substances, such as proteins and actions between them, and is based on widely accepted principles of object-oriented programming languages such as Java. Figure 3 is a diagram illustrating the object representation of molecules and relations between them. As indicated in Figure 3 there are six major classes, corresponding to the top-level classification of objects and actions: (i) a substance; (ii) a state of a substance; (iii) a similarity between substances; (iv) an action between substances; (v) a result of the action; and (vi) a mechanism that enables an action.

Figure 3 presents the class design graphically, listing the variables that represent the properties of each class or class object in the implementation. Classes can

be made nested via the mechanism of “inheritance”, *i.e.*, classes are defined starting with the most general ones and moving towards more specific classes. Definition of more specific classes is simplified because the properties of the general classes are “inherited” by the specific classes and need not be redefined each time (see, Flanagan 1997, Java in a

5 Nutshell, Second Edition. O’Reilly & Associates, Inc. Sebastopol, CA).

As shown in Figure 3, the two key object types in this scheme are substances (nodes of the graph representing regulatory networks) and actions (oriented edges connecting pairs of nodes), while result and mechanism objects are auxiliary to object action. Each substance object is characterized with a state. In this scheme, action

10 is the most complicated object; each action object is characterized by a specific pair of substances participating in the action, one of which can be active and is referred to as Subject Substance and the second of which can serve as a substrate for the former and is referred to as Object Substance. Furthermore, for each action the initial and final states corresponding to interacting substances are defined. The property Time Required of each

15 Action Object allows the setting of different durations for different actions (time is measured in relative units; see René Thomas and Richard D’Ari, 1990, "Biological Feedback," CRC Press Boca Raton, Ann Arbor, Boston).

Once developed, the specialized databases can be used to identify novel genes based on computation and analysis of phylogenetic trees for multigene families and analysis of homologous regulatory networks.

In a specific embodiment of the invention, a specialized database was generated using a set of keywords defining proteins involved in apoptosis (see, Figure 4). The specialized sequence database was referred to as Apoptosis 3. As a first step in generating the specialized database, a comprehensive set of articles describing the system of apoptosis or programmed cell death was compiled. The articles were analyzed and information on regulatory pathways characterizing apoptosis from a variety of different organisms was extracted. Such pathways included those involved in MHC-T cell receptor interactions, inflammatory cytokine signal transduction, induction by light,  $\gamma$ -radiation, hyperosmolarity or heat shock, pathways involving immunoregulatory receptors or receptors having cytoplasmic domains, integrin-related pathways and perforin/granzyme $\beta$  related pathways. The collected information was stored using Powerpoint (Microsoft) as a collection of graph/plots depicting the regulatory pathway. In addition, a list of proteins relevant to regulation of apoptosis was compiled.

Using the program Psi Retriever, sequences encoding the proteins relevant to regulation of apoptosis were retrieved from the non-redundant (NCBI) database of protein sequences and stored as a FASTA file. The FASTA file was then converted to a

binary blast database using the program FORMATDB from the BLAST suit of programs. The BLAST suit of programs provides a set of programs for very fast comparisons of a single sequence to a large database. Both the database and the search or query sequence can be any combination of nucleotide and/or amino acid sequences.

5                   In a working example described herein, the Apoptosis 3 database was used to compare genomic and cDNA sequences derived from the 13q region of human chromosome 13. This region of the chromosome is associated with Chronic Lymphocytic Leukemia (CLL). Using this method of analysis a human gene with significant homology to the mouse Rpt1 gene was identified. When the activity of Rpt1  
10 is knocked out in mice, the regulatory effect is manifested as a block in T-lymphocyte apoptosis. This result indicates that the identified human Rpt1 homology may represent the gene in which genetic defects lead to CLL.

                  The amino acid sequence of the human Rpt1 gene is presented in Figure 15. The present invention relates to nucleic acid molecules encoding the human Rpt1  
15 protein shown in Figure 15. The invention also relates to nucleic acid molecules capable of hybridizing to a nucleic acid molecule encoding the human Rpt1 protein presented in Figure 15 under conditions of high stringency. By way of example and not limitation, procedures using such conditions of high stringency are as follows: Prehybridization of filters containing DNA is carried out for 8 hours to overnight at 65°C in buffer composed



of 6x SSC, 50 mM Tris-HCl (pH7.5), 1mM EDTA, 0.02% PVP, 0.02% Ficoll, 0.02% BSA and 500 mg/ml denatured salmon sperm DNA. Filters are hybridized for 48 h at 65°C in prehybridization mixture containing 100mg/ml denatured salmon sperm DNA and 5-20 x 10<sup>6</sup> CpM of <sup>32</sup>P-labeled probe. Washing of filters is done at 37°C for 1 h in a solution containing 2x SSC, 0.01% PVP, 0.01% Ficoll and 0.01% BSA. This is followed by a wash in 0.1 x SSC at 50°C for 45 minutes before autoradiography. Other conditions of high stringency which may be used are well known in the art.

#### 5.4. GENE DISCOVERY THROUGH PHYLOGENETIC ANALYSIS OF GENE FAMILIES

The present invention provides a method for identifying novel genes comprising the following steps: (i) comparing a single sequence with a database; (ii) processing the output into a sequence alignment; (iii) computing gene trees; and (iv) analyzing the trees to predict the existence of undiscovered genes.

Figure 5 shows a “species tree,” a “gene tree” and a “reconciled tree”. A “species tree”, as defined herein, is a graph depicting the correct order of speciation events leading to a set of present day species as defined by taxonomy. A “gene tree” is a graphical representation of the evolution of a gene from a single ancestral sequence in a common progenitor to a set of present-day sequences in different species. Where gene

duplication has occurred, a branch is bifurcated. The branch lengths of a gene tree are most frequently measured either in terms of the number of amino acid or nucleotide replacements per site or in terms of millions of years (absolute geological time). In the former case, the average replacement rate in the majority of the published trees varies among tree branches, and the root-to-tip distances are different for different present day sequences. In the latter case, all root-to-tip distances are equal and the height of each interior node of the tree corresponds to the absolute geological time passed since the gene duplication corresponding to the interior node took place.

If a gene is unique, *i.e.*, represented with a single copy per genome rather than being a member of a family of similar genes, the correct gene tree depicting the origin of this gene in a few different species is identical to the species tree. In many instances, a single ancestral gene has been duplicated repeatedly during evolution to form a multigene family. A gene tree is constructed from a gene as it occurs in several species and reflects both speciation events and gene duplications within the same genome. Two homologous genes taken from different species that originated from the nearest common ancestor by speciation are referred to as orthologs, while any two genes that originated from the common ancestor via a series of events involving intragenomic duplications, or conversions, are called paralogs. The terms “ortholog” and “paralog” are applied to both nucleic acid and proteins herein.

If gene deletions are forbidden and all genes for all species represented in the tree are known, the gene tree can be reconfigured to recapitulate the species tree, such that each subtree contains only orthologous genes. This tree is referred to as a reconciled tree and is shown in Figure 5. Imperfect gene trees which contain incorrect or partial species subtrees can be used to build reconciled trees that indicate events of speciation, gene loss, and gene duplication.

Orthologs from different species in gene trees are usually clustered together, so that if all the existing homologous genes from different species were known, the same relationship of species would be recapitulated in each cluster of orthologous genes. Since in reality a considerable number of genes are not yet identified, the real gene trees contain incomplete clusters of orthologs that can be used for identification of the missing genes.

By applying phylogenetic analysis, *i.e.*, reconstruction of gene trees of gene/protein sequences, one can predict the existence of undiscovered genes in humans and other species in addition to identifying the function of a gene. Such a technique is a significantly more powerful tool for identification of new genes than mere sequence comparisons.

Methods of computing gene trees from a set of aligned sequences include the : (i) heuristic method based on an optimization principle which is not directly

motivated by a probability model (Fitch, 1974 J. Mol. Evol. 3:263-268)), (ii) the maximum likelihood method (Goldman, 1990, Syst. Zool. 30:345-361; Yang et al., 1995, Syst. Biol. 44:384-399; Felsenstein, J., 1996, Methods Enzymol. 266-418-427); and (iii) the distance matrix tree making method (Saito, N. and Nei, M., 1987, Mol. Biol. Evol. 4:406-425). Since the data analyses of orthologs and paralogs often involve very distantly related sequences, the maximum likelihood method is preferably used for small data sets and the distance-matrix method in other instances.

Sub D3  
10 To construct a reconciled tree according to the invention, the first step comprises a search for homologs in a publicly or privately available database such as, for example, GenBank, Incyte, binary BLAST databases, Swiss Prot and NCBI databases.

Following the identification of homologous sequences a global alignment is performed using, for example, the CLUSTALW program. From the sequence alignment a gene tree is constructed using, for example, the computer program CLUSTLAW which utilizes the neighbor-joining method of Saito and Nei (1997, Mol. Biol. Evol. 4:406-425).

15 Construction of a species tree is then retrieved from, for example, the following web site: <http://www.3.NCBI.NLM.NIH.GOV/taxomy.tax.html>.

The species tree and gene tree are given as input into the algorithm described below, which integrates both trees into a reconciled tree. Agreement between the gene tree and the corresponding species tree for any given set of sequences indicates

the identification of orthologs. In contrast, disagreement between the species and gene tree suggest a gene duplication that resulted in the formation of a paralog. Thus, through generation of a reconciled tree one can identify orthologs present in one species but missing in another. These can be deduced by forming subtrees of orthologs in a gene tree, and then comparing the subtree in the gene tree with a species tree. A missing gene appears as a branch present in the species tree but absent in the gene tree.

The algorithm for defining an orthologous gene subtree and predicting the undiscovered, or lost in evolution, genes is as follows:

Let  $T_g$  be the most likely gene tree identified with one of consistent tree-making methods from a set of properly aligned homologous genes  $\{1, 2, \dots, s\}$ , such that one or more homologous genes from every species corresponds to pending vertices of  $T_g$ . Each gene is labeled with the species it comes from  $(1, \dots, s)$  adding subscripts to distinguish homologous genes from the same species whenever it is necessary. Let  $T_s$  be the true species tree (tree correctly reflecting speciation events which we assume to be known) for species  $\{1, 2, \dots, s\}$ . Due to the biological meaning of  $T_s$  each species in this tree is represented only once. It is assumed that both  $T_s$  and  $T_g$  are binary, although it is straightforward to extend the algorithm described here to the case of multifurcated trees.

Algorithm

A1. For each pair of interior nodes from trees  $T_g$  and  $T_s$ , compute similarity  $\sigma(S_{gi}, S_{sj})$ .

A2. Find the maximum  $\sigma(S_{gi}, S_{sj})$ .

A3. Save  $S_{gi}$  as a new subtree of orthologs, save  $\{S_{gi}\} - \{S_{sj}\}$  as a set of species that  
5 are likely to have gene of this kind (or lost it in evolution).

A4. Eliminate  $S_{gi}$  from  $T_g$ ;  $T_g := T_g \setminus S_{gi}$ .

A5. Continue A2 - A4 until  $T_g$  is non-empty.

The following definitions apply:

Let  $S_{gi}$  be an  $i$ th subtree of  $T_g$  (corresponding to the  $i$ th interior node), correspondingly, let  
10  $S_{sj}$  be  $j$ th subtree of tree  $T_s$ .

Let  $\{S_{gi}\}$  stand for an unordered set of species represented in  $S_{gi}$  such that each species is represented exactly once, and let  $|\{S_{gi}\}|$  and  $|\{S_{gi}\}|$  be the number of entries in  $\{S_{gi}\}$  and the number of pending vertices in  $S_{gi}$ , respectively. Define by  $S_{sj}(S_{gi})$  the unique subtree of  $S_{sj}$  that has leaves labeled exclusively with species from  $|\{S_{gi}\}|$ , so that each  
15 element of  $|\{S_{gi}\}|$  is used i.e., that is, the unique subtree obtained by eliminating from  $S_{sj}$  all species that are not present in  $|\{S_{gi}\}|$ .

Then define similarity measure,  $\sigma$ , between  $S_{gi}$  and  $S_{sj}$  in the following way:

$\sigma(S_{gi}, S_{sj}) = 0$  if  $|S_{gi}| \neq |\{S_{gi}\}|$ , or  $S_{sj}(S_{gi}) \neq S_{gi}$ , and

$$\sigma(S_{gi}, S_{sj}) = |S_{gi}|$$

The support of tree clusters by data can be measured using the bootstrap technique described in Felsenstein (1985, Evolution 39:783-791).

In an embodiment of the invention, the human antiquitin gene was identified using phylogenetic analysis. The aldehyde dehydrogenase gene family in humans can be subdivided into at least ten ancient subtrees characterized by different functions of corresponding proteins. These genes probably arose from a series of gene duplications of an ancestral gene which took place before the divergence of a common ancestor of Eukaryotes and Eubacteria.

The aldehyde dehydrogenase gene cluster is highlighted in Figure 6 which shows the original tree of ALDH sequences, the circled area indicating a sequence cluster where bacterial (*Bacillus subtilis*), plant (*Brassica napus*), and nematode (*Caenorhabditis elegans*) ortholog is present, but a human ortholog is not known. A random screening of cDNA libraries showed that a human ortholog, referred to as antiquitin, does exist. Figure 7 shows the same gene tree as in Figure 6 with an additional human protein referred to as antiquitin present in the tree.

In yet another embodiment of the invention, a human ortholog of the murine Max-interacting transcriptional repressor Mad3 was identified through phylogenetic analysis of a gene family. The gene tree was constructed as follows. The protein sequences of known members of the *Mad* gene family were extracted from

GenBank database. The extracted sequences were aligned using multiple alignment program CLUSTALW running on Sun SPARC station. Redundant and non-homologous sequences as well as distant homologs from *S. cerevisiae*, *C. elegans*, *D. melanogaster* etc. were removed from the alignment. The refined set of sequences were realigned with CLUSTALW and a gene tree as presented in Figure 18A was computed. To identify a human ortholog of the Mad3 protein, a human dbEST at NCBI was searched with program TBLASTN using mouse Mad3 protein sequences as a query. Two highly homologous ESTs were identified and are presented in Figure 17A. To obtain a complete coding sequence a search was conducted to obtain overlapping sequences in dbEST. The search for overlapping sequences was performed using the program Iterate with EST Zs77e55.rl (gb/AA278224) as the search query. The search identified a single overlapping sequence. The search for overlapping sequences was performed using program Iterate with EST zs77e55.rl (gb/AA278224) serving as a query. The search returned a single overlapping sequence, namely HUMGS0012279 (dbj/C02407), thus showing that the two EST sequences found during the initial TBLASTIN search belong to the same gene. The complete sequence of the gene was assembled from the two ESTs using commercially available sequence assembly program SeqMan11 (DNASTAR Inc., WI). The nucleotide sequence of the human Mad3 gene is presented in Figure 17B. The



deduced amino acid sequence of which is presented in Figure 17C. The complete DNA sequence is also shown.

The present invention relates to nucleic acid molecules encoding the human Mad3 protein shown in Figure 17C. The invention also relates to nucleic acid molecules that hybridize to the nucleic acid molecule of Figure 17B under conditions of high stringency and encode a Mad3 protein. By way of example and not limitation, procedures using such conditions of high stringency are as follows: Prehybridization of filters containing DNA is carried out for 8 hours to overnight at 65°C in buffer composed of 6x SSC, 50mM Tris-HCl (pH7.5), 1mM EDTA, 0.02% PVP, 0.02% Ficoll, 0.02% BSA and 500 mg/ml denatured salmon sperm DNA. Filters are hybridized for 48 hours at 65°C in prehybridization mixture containing 100 mg/ml denatured salmon sperm DNA and 5-20 x 10<sup>6</sup> CpM of <sup>32</sup>P-labeled probe. Washing of filters is done at 37°C for 1 hour in a solution containing 2x SSC, 0.01% PVP, 0.01% Ficoll and 0.01% BSA. This is followed by a wash in 0.1x SSC at 50°C for 45 minutes before autoradiography. Other conditions of high stringency which may be used are well known in the art.

### 5.5. SIMULATION AND HYPOTHESIS TESTING

The simulation and hypothesis testing methods of the invention, described in the subsections below, utilize specialized databases of gene/protein structures and

interactions for identifying potentially undiscovered members of multigene families through comparisons of regulatory networks for different species, searching expressed sequence tag (EST) databases, and simulation of regulatory cascades.

#### 5.5.1. GENE DISCOVERY THROUGH ANALYSIS OF REGULATORY NETWORKS

5

The present invention provides a method for identifying undiscovered genes through comparisons of regulatory networks for different species where functionally similar regulatory systems are conserved. The amount of information available concerning regulatory genes and/or proteins in different organisms and their functional relationships allows one to reconstruct and compare regulatory networks. Since in most cases, the knowledge of all genes involved in almost any particular regulatory system is incomplete, a comparison of homologous networks within the same organism and between different species permits the identification of genes absent in a system under comparison.

10

15

The identified genes, being part of a regulatory network, are implicated as potentially contributing to a phenotype of a disease associated with the system under analysis. Using the methods of the present invention these putative disease genes can be

cloned, mapped and analyzed for mutations directly, thereby omitting the expensive and time-consuming steps of positional cloning and sequencing of genomic regions.

Gene discovery by analysis of regulatory networks is outlined in Figure 8. The analysis is initiated starting with a biological system (*e.g.*, signaling pathway of genes involved in Bcl-2-regulated apoptosis in lymphocytes), a single gene (*e.g.*, Bcl-2) or a gene family (*e.g.*, caspases).

Initially, a specialized database is generated for comparison of regulatory networks between different species. For example, starting with a single candidate gene in a single species, a typical iteration in this process begins with identification of all known proteins and genes that are upstream and downstream with respect to it in regulatory hierarchies and the reconstruction of a network of interacting genes and proteins. Next, for each protein, a set of key domains and motifs is identified and this information is used to search for related proteins in humans and other species. The identified sequences are compared and for each pair of sequences showing similarity above a certain threshold, a similarity object is generated. A similarity object is generated if two sequences, nucleotide or amino acid, show significant similarity in database searches ( $p$  value  $< 0.001$ ). The object retains the following information: (i) reference to similar substances *i.e.*, genes or proteins; (ii) significance of the similarity, similarity score and percent of identity; and (iii) coordinates of the similarity region within two compared sequences.

“Orthology objects” constitute a subset of “similarity objects” which satisfies one additional requirement, *i.e.*, that two similar sequences should be identified as orthologs by the tree-based algorithm described above. In identifying orthologs, if gene A is orthologous to gene B, and gene B is orthologous to gene C, gene A is necessarily  
 5 orthologous to gene C.

In a specific embodiment of the invention, for each species under analysis, orthologous proteins or genes are identified. In a further embodiment of the invention, small orthologous molecules participating in a regulatory network for two or more species may also be identified. Where proteins, genes, or molecules are orthologs, the  
 10 action of the protein, gene or molecule between species may be interchangeable. If more than two species are involved in the analysis, subtrees of orthologous substances and subtrees of orthologous actions are identified.

Once orthologous genes, proteins or molecules are identified in two or more species, by forming a reconciled tree, for example, a set of orthologous or  
 15 paralogous regulatory networks can be analyzed and visualized using graph theory where arcs represent actions and vertices represent substances. Thus, the method of the invention may further comprise the following steps: (i) superimposing the orthologous regulatory networks from two or more species and searching for the actions (arcs) and substances (vertices) in the homologous networks that are represented in some taxa but

absent in others; (ii) superimposing paralogous regulatory networks from the same taxa and searching for paralogous genes that are missing in some taxa; and (iii) computing a general regulatory network that summarizes common regulatory sequence relationships known for more than one species.

5                   In a specific embodiment of the invention a set of regulatory networks from different species, relating to the same biological system, apoptosis, for example, can be analyzed and visualized utilizing the following methods: (i) for each species functional information is collected relating to apoptosis; (ii) using the functional information, regulatory networks for each species comprised of interacting proteins and/or the genes  
10 involved in apoptosis are generated; (iii) the sequences of the interacting proteins and genes of each of the regulatory network are compared and for sequences showing similarity above a predetermined threshold range; and (iv) distinguishing between orthologs and paralogs using the methods set forth above.

                  An analysis similar to that performed using subtrees of sequences may be  
15 applied to classify protein functions as orthologous or paralogous actions. A “generalized” regulatory network maybe represented as a network wherein a substance as it occurs in a particular species is substituted with a cluster (i.e., subtree) of orthologous substances among species. In the final step of the analysis the clusters within each species are compared to one another, to identify missing genes.

Figure 11 depicts the regulatory relationships among hypothetical proteins (denoted with Arabic numerals) of hypothetical species A and B. As indicated in Figure 11A, an overlay of regulatory data for two species overlaps, but not completely. As indicated, protein 5 is known only for species B while protein 3 is known only for species

- 5 A. The proteins in different species denoted with the same numeral are considered orthologous. As indicated, the regulatory relationships between a pair of proteins can be of three different kinds. Figure 9B, 9C, and 9D represent Boolean operations, OR, AND, and XOR, as arcs of the two regulatory relationships depicted in Figure 9A, the same operations being applicable to the set of vertices of the two regulatory relationships.
- 10 In some instances, orthologous networks in two distantly related taxa may have the same domains but arrangement of the domains between the related taxa may be different. In such a case, a one-to-one correspondence between orthologous proteins in closely related species has to be substituted with a one-to-many relationship among domains comprised within the proteins. For this purpose, a similarity object may be defined operating on
- 15 pairs of motifs/domains in two proteins, and substitute pairs of orthologous proteins with pairs of orthologous domains. After this correction, homologous networks are compared as described above.

Figure 10 is a diagram representing a hypothetical example of defining homologous protein networks in two different species using protein motifs, the diagram

showing only two hypothetical proteins (lane 2) for species A and three hypothetical proteins (lanes 1, 3, and 4) for species B. Protein 1 in both species has motifs  $\alpha$  and  $\beta$ , protein 2 has motifs  $\delta$ ,  $\epsilon$ , and  $\zeta$ , and proteins 3 and 4 have motifs  $\delta$  and  $\zeta$ , and  $\epsilon$ , respectively. The motif analysis indicates that proteins 3 and 4 in species B may collectively perform the same function as protein 2 in species A.

#### 5.5.2 GENE DISCOVERY BASED ON PROTEIN MOTIF/DOMAIN SEARCHES

The present invention provides yet another method for identifying genes that are homologous and perform the same or an analogous function in different species.

10 The method of the invention comprises the following steps: (i) creating a database of sequences which comprise a motif or domain composition of a gene of interest using, for example, HMMER software; and (ii) searching additional databases for expressed sequence tags (ESTs) containing the domains and motifs characteristic for the gene of interest with HMMs of domains and motifs identified in step (i). In yet another

15 embodiment of the invention, sequences may be searched which correspond to nucleotide sequences in an EST database or other cDNA databases using a program such as BLAST and retrieving the identified sequences. In an optional step, for each EST identified, sequence databases can be searched for overlapping sequences for the purpose of assembling longer overlapping stretches of DNA. Once identified, the ESTs can be used

to isolate full length nucleotide sequences comprising the gene of interest using methods such as those described in Section 5.4, *infra*.

The general flowchart scheme for gene discovery analysis based on motif/domain search is shown in Figure 11. In a specific embodiment of the invention,

5 the method referred to as the "phylogenetic reflection technique" comprises, first, defining the motif or domain composition of a gene of interest involved in a biological system of interest. Second, protein-coding genes from other species, including for example yeast and/or nematode genes, that bear a significant similarity to the gene of interest or a specified domain of the corresponding protein are collected. Third, the identified genes

10 are in turn subjected to a "domain analysis" to establish protein motifs which might suggest a function of these genes using, for example, HMMER software. Fourth, the selected genes are in turn used for database searches in EST databases (dbEST) and/or a non-redundant (nr) database to identify unknown genes that are potentially orthologous to the selected yeast and nematode genes. Once identified ESTs having different tumor

15 suppressor domains may be linked using multiple PCR primers. Using routine cloning techniques, well known to those of skill in the art, a full length cDNA representing the gene of interest can be obtained.

Once new genes are identified by domain/motif analysis experimental searches may be carried out to isolate complete coding sequences and evaluate their



tissue- and disease-specific expression patterns. In parallel their position with respect to regulatory networks can be identified as described below.

In a specific embodiment of the invention, an apoptosis related human gene was identified using the method described above. As a first step *C. elegans* genes containing either POZ or Kelch domains were identified. A Hidden Markov Model was developed using POZ and Kelch sequences from the *Drosophila* Kelch protein and any identified homologs. The resulting Hidden Marker Model was used to search through the collection of *C. elegans* protein sequences. One of the identified *C. elegans* genes contained a POZ domain, death domain, kinase domain and heat repeat. The presence of both a death domain and a kinase domain suggested that the protein functions as a regulatory protein.

A human EST database was searched using the protein sequence of the identified *C. elegans* gene and two sequences were identified (Figure 14A). A gene tree was computed to determine whether the identified human sequences were orthologs of the *C. elegans* gene. As depicted in Figure 14B, the human EST AA481214 appears to be a true ortholog of the *C. elegans* gene. Figure 14C presents the nucleotide sequence of the identified death domain gene. Figure 14D presents the amino acid sequence of the death domain protein.

The present invention encompasses the nucleic acid molecule of Figure 14C, comprising the sequence of EST AA481214 and proteins encoded by said nucleic acid molecule. The invention also relates to nucleic acid molecules capable of hybridizing to such a nucleic acid molecule under conditions of high stringency. By way of example and not limitation, procedures using such conditions of high stringency are as follows: Prehybridization of filters containing DNA is carried out for 8 hours to overnight at 65°C in buffer composed of 6x SSC, 50mM Tris-HCl (pH7.5), 1mM EDTA, 0.02% PVP, 0.02% Ficoll, 0.02% BSA and 500 mg/ml denatured salmon sperm DNA. Filters are hybridized for 48 hours at 65°C in prehybridization mixture containing 100 mg/ml denatured salmon sperm DNA and 5-20 x 10<sup>6</sup> CpM of <sup>32</sup>P-labeled probe. Washing of filters is done at 37°C for 1 hour in a solution containing 2x SSC, 0.01% PVP, 0.01% Ficoll and 0.01% BSA. This is followed by a wash in 0.1x SSC at 50°C for 45 minutes before autoradiography. Other conditions of high stringency which may be used are well known in the art.

#### 5.5.3. SIMULATION OF REGULATORY CASCADES

In an embodiment of the invention, an interactive graphical program is utilized for visualizing the scheme of regulatory relationships, "current" states of the

substances, and active and inactive actions between pairs of substances. Such a program can be utilized for identification of genes which are associated with a specific disease.

Currently, disease associated genes are discovered through positional cloning methods which combine methods of genetics and physical mapping with mutational analysis. The

5 present invention provides a novel method for discovering disease associated genes.

For simulating regulatory cascades, it is assumed that the time in a simulated regulatory system advances in discrete "quanta," or periods of time. The "state of substances" of

the system for each discrete period of time is computed by: creating a set of substance

objects, where a set of interactions between each created substance object is known, an

10 initial state is specified. The time is initially set to zero. All defined actions are observed

to confirm that the substances corresponding to the actions (i) exist, and (ii) are in the

right initial states. Action is defined by a pair of substances that are in suitable states.

The "subject" substance is in the inactive state, while the "object" substance can be in

either active, or inactive, state depending on the action type. For example, the action

15 "dephosphorylation" requires an active phosphatase ("subject" substance) and a

phosphorylated substitute protein ("object" substance) in phosphorylated form. If both

conditions are satisfied, the action is recorded as in progress. At termination, the

substances must change their states as specified by the action. On each following

"quantum" of time, the simulation proceeds in the same way while maintaining the

“bookkeeping” of the remaining time for each action and the remaining lifespan of each substance. The simulation stops when there are no more active actions available. The program allows editing of the properties of the objects, changing the scale and focus of the visualized simulation, and experimenting with the systems output.

5                   In a specific embodiment of the invention a “knock out” of a gene can be simulated to model the regulatory system that normally includes hypothetical gene A. One of the typical questions related to the gene knock out is how does the knock out affect a biological pathway of interest. A hypothetical example of evaluating the impact of a knock out of hypothetical gene A on the expression of a hypothetical gene B is  
10   shown in Figure 12. The answer to such a question could be “gene B will be inhibited” or “gene B will be induced” or “no effect”.

                  In the practice of the present invention, a simple algorithm involving multiplication of gene interaction “signs” along the shortest pathway between the genes can be used to determine the outcome. The algorithm involves the following steps: (i)  
15   identification of the shortest non-oriented pathway connecting genes A and B involved in a pathway of interest; (ii) assigning sign “-” to gene A since it is knocked out and taking this sign as the initial sign value; (iii) moving along the shortest pathway between genes A and B, multiplying the current value of the sign with the sign of the next arc, where “-” stands for inhibition, “+” stands for induction or activation, and “0” stands for the lack of

interaction between two proteins in the specified direction; (iv) determining if the final result of multiplication is "0", if so eliminating the zero arc and trying to find the shortest oriented bypass pathway between A and B in the remaining network; otherwise stop.

The final value of the sign at the moment of arriving at vertex B would indicate the most likely effect of the knock out of gene A which can be any one of the following: inhibition of gene B, induction/activation of gene B, or none. In addition to the "electronic knock out", an "electronic knock in" of a particular gene can be simulated. In such a computer simulation, the artificial addition of a gene and its effect on a regulatory system may be analyzed.

#### 5.6. IDENTIFICATION AND ISOLATION OF NOVEL GENES

The present invention relates to identification of novel genes, i.e., missing orthologs or paralogs, and the isolation of nucleic acid molecules encoding novel genes.

In a specific embodiment, a nucleic acid molecule encoding a missing ortholog or

paralog can be isolated using procedures well known to those skilled in the art (See, for example, Sambrook et al., 1989, Molecular Cloning, A Laboratory Manual, 2d Ed., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York Glover, D.M. (ed.), 1985, DNA Cloning: A Practical Approach MRL Press, Ltd., Oxford, U.K. Vol. I, II.).

For example, genomic and/or cDNA libraries may be screened with labeled DNA fragments derived from a known ortholog or paralog from a specific species and hybridized to the genomic or cDNA libraries generated from a different species. For cross species hybridization, low stringency conditions are preferred. For same species hybridization, moderately stringent conditions are preferred. Any eukaryotic cell potentially can serve as the nucleic acid source for the molecular cloning of the gene of interest. The DNA may be obtained by standard procedures known in the art from cloned DNA (e.g., a DNA "library"), by cDNA cloning, or by the cloning of genomic DNA, or fragments thereof, purified from the desired cell.

By way of example and not limitation, procedures using conditions of low stringency are as follows (see also Shilo and Weinberg, 1981, Proc. Natl. Acad. Sci. USA 78:6789-6792; and Sambrook et al. 1989, Molecular Cloning, A Laboratory Manual, 2d Ed., Cold Spring Harbor Laboratory Press, Cold Spring harbor, New York): Filters containing DNA are pretreated for 6 h at 40°C in a solution containing 35% formamide, 5X SSC, 50 mM Tris-HCl (pH 7.5), 5 mM EDTA, 0.1% PVP, 0.1% Ficoll, 1% BSA, and 500 mg/ml denatured salmon sperm DNA. Hybridizations are carried out in the same solution with the following modifications: 0.02% PVP, 0.02% Ficoll, 0.2% BSA, 100 mg/ml salmon sperm DNA, 10% (wt/vol) dextran sulfate, and 5-20 X 10<sup>6</sup> cpm <sup>32</sup>P-labeled probe is used. Filters are incubated in hybridization mixture for 18-20 h at 40°C, and

then washed for 1.5 h at 55°C in a solution containing 2X SSC, 25 mM Tris-HCl (pH 7.4), 5 mM EDTA, and 0.1% SDS. The wash solution is replaced with fresh solution and incubated an additional 1.5 h at 60°C. Filters are blotted dry and exposed for autoradiography. If necessary, filters are washed for a third time at 65-68°C and reexposed to film. Other conditions of low stringency which may be used are well known in the art (*e.g.*, as employed for cross species hybridizations).

In another specific embodiment, a nucleic acid which is hybridizable to a nucleic acid under conditions of moderate stringency is provided. For example, but not by way of limitation, procedures using such conditions of moderate stringency are as follows: filters containing DNA are pretreated for 6 h at 55°C in a solution containing 6X SSC, 5X Denhart's solution, 0.5% SDS and 100 mg/ml denatured salmon sperm DNA. Hybridizations are carried out in the same solution and 5-20 X 10<sup>6</sup> CpM <sup>32</sup>P- labeled probe is used. Filters are incubated in the hybridization mixture for 18-20 h at 55°C, and then washed twice for 30 minutes at 60°C in a solution containing 1X SSC and 0.1% SDS. Filters are blotted dry and exposed for autoradiography. Other conditions of moderate stringency which may be used are well-known in the art. Washing of filters is done at 37°C for 1 h in a solution containing 2X SSC, 0.1% SDS.

For expression cloning (a technique commonly used in the art), an expression library is constructed. For example, mRNA is isolated from the cell type of

interest, cDNA is made and ligated into an expression vector (*e.g.*, a bacteriophage derivative) such that it is capable of being expressed by a host cell (*e.g.*, a bacterium) into which it is then introduced. Various screening assays can then be used to select for the expressed gene product of interest based on the physical, chemical, or immunological properties of its expressed product. Such properties can be deduced from the properties of the corresponding orthologs from other species.

In another embodiment, polymerase chain reaction (PCR) can be used to amplify the desired sequence from a genomic or cDNA library. To isolate orthologous or paralogous genes from other species, one synthesizes several different degenerate primers, for use in PCR reactions. In a preferred aspect, the oligonucleotide primers represent at least part of the gene comprising known ortholog or paralog sequences of different species. It is also possible to vary the stringency of hybridization conditions used in priming the PCR reactions, to allow for greater or lesser degrees of nucleotide sequence similarity between the known nucleotide sequences and the nucleic acid homolog being isolated.

Synthetic oligonucleotides may be utilized as primers to amplify by PCR sequences from a source (RNA or DNA), preferably a cDNA library, of potential interest. PCR can be carried out, *e.g.*, by use of a Perkin-Elmer Cetus thermal cycler and a thermostable polymerase, *e.g.*, Amplitaq (Perkin-Elmer). The nucleic acids being



amplified can include mRNA or cDNA or genomic DNA from any eukaryotic species.

After successful amplification of a segment of a the gene of interest, that segment may be molecularly cloned and sequenced, and utilized as a probe to isolate a complete cDNA or genomic clone.

5                   Once identified and isolated the gene of interest can then be inserted into an appropriate cloning vector for amplification and/or expression in a host. A large number of vector-host systems known in the art may be used. Possible vectors include, but are not limited to, plasmids and modified viruses, but the vector system must be compatible with the host cell used. Such vectors include, but are not limited to,  
10 bacteriophages such as lambda derivatives, or plasmids such as pBR322 or pUC plasmid derivatives or the Bluescript vector (Stratagene). The insertion into a cloning vector can, for example, be accomplished by ligating the DNA fragment into a cloning vector which has complementary cohesive termini.

15                   6. EXAMPLE: USE OF SPECIALIZED DATABASES  
                    FOR IDENTIFICATION OF NOVEL GENES

To test the method of using databases for gene discovery, protein sequence and domain/motif databases specific to two overlapping functional groupings of proteins:

(i) proteins known to be tumor suppressors, and (ii) proteins implicated in apoptosis in  
20 animals were developed.

## 6.1 APOPTOSIS GENE DISCOVERY METHOD

Identification of a putative apoptosis-related human gene began with an identification of all genes in *C. elegans* that contained either a POZ or kelch domain. A subset of these genes is shown in Figure 13. Hidden Markov Models (HMM) for the POZ and Kelch domains were built as follows. Starting with POZ and kelch sequences from the *Drosophila* kelch protein (gi|577275) homologs were identified in other protein sequences using the BLASTP program. The resulting sequences showing significant similarity (e-value less than 0.001) were aligned using CLUSTALW program and the alignments were used to build Hidden Markov Models with HMMER-2 package (Krogh et al., 1995, :<http://hmmer.wustl.edu/>). A computer printout listing of HMM models of tumor suppressors appears as a Microfiche H to the present specification. (See, <http://hmmer.wustl.edu>; Chapter 2, which is incorporated by reference herein in its entirety, for a detailed description of HMM models)

The resulting models were used to search through a database collection of *C.elegans* protein sequences. The domain structures of proteins having either a POZ or kelch domain were identified using existing collections of protein domains (e.g., see [http://blocks.fhcrc.org/blocks/blocks release.html](http://blocks.fhcrc.org/blocks/blocks%20release.html), <http://coot.embl-heidelberg.de/SMART/>, <http://www.motif.genome.ad.jp/>).

One of the unannotated protein-coding genes of *C. elegans* (corresponding protein accession number gi|1132541, see Figure 11) appeared to include a POZ domain, death domain, kinase domain, and heat repeat. A death domain is characteristic for the apoptosis system and a kinase domain indicates that the protein is likely to participate in phosphorylation of other proteins. The presence of these particular domains suggests that this protein is serving as a regulatory protein.

Using the protein sequence of gi|1132541, the database of human EST sequences was searched and a number of partial human cDNA sequences representing potential human orthologs or paralogs of the *C.elegans* gi|1132541 were identified.

- 10 The two closest human sequences, AA481214 and W51957, are depicted in Figure 14A. To determine whether the identified human sequences were orthologs or paralogs to the gi|1132541 gene of *C. elegans*, a gene tree (Saito and Nei, 1997, Molecular Biol. Evol. 4:406-425) was computed. The gene tree was generated using homologous genes identified with a BLASTP search against NCBI non-redundant database, using the human
- 15 EST AA481214 sequence as a query. The resulting tree indicates that the identified human EST AA481214 represents a true ortholog of the *C.elegans* gene gi|1132541 (Figure 14B). The nucleotide sequence of the death domain protein is shown in Figure 14C, as well as the deduced amino acid sequence presented in Figure 14D.

### 6.1.2 APOPTOSIS GENE DISCOVERY METHOD

As a first step in identifying a novel gene involved in apoptosis, a comprehensive set of articles describing the system of apoptosis/programmed cell death in different species was compiled using the keyword "apoptosis". By analyzing the articles, information on regulatory pathways characterizing this system in different species, *i.e.*, *C. elegans*, mouse, fruit fly, chicken, and human, was extracted. The regulatory information was stored as a collection of schemes produced in PowerPoint (Microsoft). Figure 4 shows a set of keywords defining proteins involved in apoptosis pathways. The keywords were used to generate a specialized sequence database, referred to as Apoptosis3, utilizing the PsiRetriever program for extraction of proteins from the all-inclusive non-redundant GenBank database (NCBI). Using program PsiRetriever, sequences from the non-redundant (NCBI) database of protein sequences, were retrieved and stored as a FASTA file. The FASTA file was then converted into binary blast database using program FORMATDB from the BLAST suit of programs.

Genomic and cDNA sequences located in the region of human chromosome 13q were compared with the Apoptosis3 database using BLASTALL program from BLAST program complex. This region of the human genome is associated with Chronic Lymphocytic Leukemia (CLL). The comparison revealed significant similarity between a CLL region open reading frame and the mouse RPT1 protein

(sp|P15533|RPT1) (Figure 13). Analysis of regulatory functions of RPT1 in the mouse reveals that this gene functions as a repressor of the interleukin 2 receptor (IL-2R) gene. When the RPT1 gene is knocked out, the regulatory effect is manifested as a block of the apoptotic pathway in T lymphocytes resulting in an accumulation of T lymphocytes in blood. This result is consistent with aberrations observed in CLL, namely abnormal accumulation of B-cells in the blood ( Trentin L. et al., 1997, Leuk. Lymphoma 27:35-42) and mutations in the human RPT1 gene play a role in development of CLL.

#### 6.1.3 EXAMPLE: A DISCOVERY OF A HUMAN ORTHOLOG OF THE MURINE MAX-INTERACTING TRANSCRIPTIONAL REPRESSOR

The family of *Myc* proto-oncogenes encodes a set of transcription factors implicated in regulation of cell proliferation, differentiation, transformation and apoptosis. C-*Myc* null mutations result in retarded growth and development of mouse embryos and are lethal by 9-10 day of gestation. In contrast, overexpression of *Myc* genes inhibits cell differentiation and leads to neoplastic transformation. Moreover, deregulation of *Myc* expression by retroviral transduction, chromosomal translocation or gene amplification is linked to a broad range of naturally occurring tumors in humans and other species.

Another protein, called *Max*, is an obligatory heterodimeric partner for *Myc* proteins in mediating their function as activators of transcription during cell cycle progression, neoplastic transformation and programmed cell death (apoptosis). In order to make an active transcription factor the *Myc* proteins must form heterodimers with *Max* protein. This interaction with *Max* protein is necessary for specific binding of *Myc* with CACGTG box (or related E-boxes) on DNA and for activation of promoters located proximal to the binding sites.

Besides the *Myc* family of transcription factors, the *Max* protein forms complexes with another family of so-called *MAD* proteins: *Mxi1*, *MAD1*, *MAD3* and *MAD4*. Whereas *Myc:Max* complexes activate transcription, *MAD:Max* complexes work in an opposite way repressing the transcription through the same E-box binding sites and apparently antagonize *Myc*-mediated activation of the same set of target genes.

During tissue development a shift from *Myc:Max* to *MAD:Max* complexes occurs coincidentally with the switch from cell proliferation to differentiation. The switch in heterocomplexes is thought to reflect a switch from activation to repression of common genes leading to cessation of proliferation, exiting the cell cycle and the beginning of cell differentiation. In differentiating neurons, primary keratinocytes, myeloid cell lines and probably other tissues the expression of different *MAD:Max* complexes appear in sequential order during the transition from cell proliferation to differentiation. The *MAD3*

expression appears first and it is restricted to proliferating cells prior to differentiation where it is co-expressed with two different member of *Myc* family, c-*Myc* or N-*Myc*. *Mxi1* transcripts are detected in proliferating and differentiating cells whereas *MAD1* and *MAD4* were confined to post-mitotic cells. Because *Myc* expression is not always

5 downregulated in post-mitotic cells, co-expression of *Myc* and *MAD* genes may result in competition for *Max* heterodimers thus providing promoting or inhibitory effect on cell proliferation.

The gene expression patterns, along with ability of Mad proteins to suppress *Myc*-dependent transformation, are consistent with a potential function of Mad

10 genes as tumor suppressors. This view is supported by the fact that allelic loss and mutations were detected at the *Mxi1* locus in prostate cancers (Eagle et al., 1995 Nat Genet 9:249-55). Cloning of the murine proteins *Mad3* and *Mad4* as well as their relation to *Max* signaling network was described by Hurlin (Hurlin PJ, et al., 1995, EMBO J. 14:5646-59) and Queva (Queva et al. 1998 Oncogene 16:967-977). Human

15 orthologs of *Mad4*, *Mad1* and *Mxi1* are known.

In this example, the discovery of an unknown human ortholog of *Mad3* protein found “*in silico*,” by means of phylogenetic analysis of known mouse and human members of the *Mad* gene family and database searches is described. Since the function

of murine *Mad3* as a *Max*-interacting transcriptional repressor of *Myc*-induced neoplastic transformation is well described, we can assign the same function to its human ortholog. The gene tree shown in the Figure 20 was constructed in the following way. The protein sequences of known members of *Mad* gene family were extracted from GenBank

5 database using NCBI Entrez keyword searches. The extracted sequences were aligned using multiple alignment program Clustalw running on Sun SPARC station. The quality of the multiple alignment was checked using program HitViewer Iterate (A. Rzhetsky, available upon request) and the redundant, non-homologous sequences as well as distant homologs from *S. cerevisiae*, *C. elegans*, *D. melanogaster* etc. were removed from the  
10 alignment. The refined set of sequences was realigned with Clustalw and a gene tree as presented in Figure 15A was computed from the alignment using program NJBOOT ([http://genome6.cpmc.columbia.edu // andrey](http://genome6.cpmc.columbia.edu//andrey)) running on Sun SPARC station and viewed with program TreeView ([http://genome6.cpmc.columbia.edu // andrey](http://genome6.cpmc.columbia.edu//andrey)).

The tree presented in Fig.19A clearly shows the relationships between  
15 three known mouse genes and their two human homologs. Attempts to find a missing human ortholog of the mouse *Mad3* gene in protein non-redundant database at NCBI using BLAST search did not identify any human homologs other than sequences that were already present on the tree, confirming the absence of a known human ortholog for *Mad3* protein in the database.



In order to identify a human ortholog of the Mad3 protein, a human dbEST at NCBI was searched with program TBLASTN using Mad3 protein sequence as a query. Two EST were identified and are shown in Figure 17A.

Due to the nature of dbEST database this search produced only partial  
 5 sequences of potential candidate genes. To obtain complete coding sequences (complete cds) of the genes, a search was conducted to obtain overlapping sequences in dbEST. The search for overlapping sequences was performed using the program Iterate with EST zs77e55.r1 (gb|AA278224) serving as a query. The search returned a single overlapping sequence, namely HUMGS0012279 (dbj|C02407), thus indicating that the two EST  
 10 sequences found during the initial TBLASTN search belong to the same gene.

The complete sequence of the gene was assembled from the two ESTs using commercially available sequence assembly program SeqManII (DNASTAR Inc., WI). The nucleotide sequence of the human *Mad3* gene is presented in Figure 17B. The deduced amino acid sequence of the gene is presented in Figure 17C. The translated sequence consists  
 15 of 206 amino acid residues 81% of which are identical to mouse Mad3 protein. The alignment of human and mouse Mad3 proteins shown below was made using BLAST server at NCBI and is presented in Figure 17C.

Multiple alignment of the new sequence with sequences of known Mad proteins was made using Clustalw and viewed with the HitViewer. A gene tree was

computed from this alignment using NJBOOT. Multiple alignment of the new sequence with sequences of known Mad proteins (Figure 17C) along with its position on gene tree (Figure 18B) shows that this new human gene found by the approach described above belongs to the family of Mad proteins and is the ortholog of mouse Mad3.

5           The present invention is not to be limited in scope by the specific embodiments described herein, which are intended as single illustrations of individual aspects of the invention, and functionally equivalent methods and components are within the scope of the invention. Indeed, various modifications of the invention, in addition to those shown and described herein will become apparent to those skilled in the art from  
10 the foregoing description and accompanying drawings. Such modifications are intended to fall within the scope of the appended claims.

Various publications are cited herein, the contents of which are hereby incorporated by reference in their entireties.